ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2.
Available from http://www.i6doc.com/en/.

Online Adaptation of Compressed Models by Pre-Training and Task-Relevant Pruning

Thomas Avé[1], Matthias Hutsebaut-Buysse[1], Wei Wei[1] and Kevin Mets[2] *

University of Antwerp - imec, Sint-Pietersvliet 7, 2000 Antwerp, Belgium
IDLab - Department of Computer Science [1], Faculty of Applied Engineering[2]

Abstract. Neural networks are increasingly deployed on edge devices, where they must adapt to new data in dynamic environments. Here, model compression techniques like pruning are essential. This involves removing redundant neurons, increasing efficiency at the cost of accuracy, and creating a conflict between efficiency and adaptability. We propose a novel method for training and compressing models that maintains and extends their ability to generalize to new data, improving online adaptation without reducing compression rates. By pre-training the model on additional knowledge and identifying the parts of the deep neural network that actually encode task-relevant knowledge, we can effectively prune the model by 80% and achieve 16% higher accuracies when adapting to new domains.

1 Introduction

Deep learning has surged in popularity over the past decade, with swarm intelligence as an emerging application domain where low-power sensor devices coordinate to complete complex tasks. Online tuning by individual nodes is crucial in dynamic environments with varying local conditions. However, limited on-device resources constrain the complexity of deployable and trainable models. Model compression addresses this by reducing neural network size while preserving predictive power. Pruning, one such method, removes redundant parameters from a fully-trained network, resulting in a more compact architecture. However, traditional methods do not account for online learning scenarios, focusing instead on creating the smallest static model, removing any redundancy not necessary for the current task, including features beneficial for new tasks.

We propose a novel method that extends the generalization capabilities of compressed models, enabling online adaptation without sacrificing compression rates. Generalization involves learning general patterns that apply to new data, which is crucial for adapting to changing data distributions. Prior research suggests that pre-training on a larger dataset before fine-tuning improves generalization [2] but this typically requires larger models that are harder to prune due to more active connections encoding additional knowledge [1]. Methods like Iterative Magnitude Pruning (IMP) use weight magnitudes to determine importance, making it challenging to distinguish between task-relevant and generalization-improving knowledge. This limits the compression potential, making it counterproductive for low-power online adaptation.

*This paper was supported by the OpenSwarm project and FWO [Grant 1SD9523N].

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

To address this issue, we leverage Layer-wise Relevance Propagation (LRP), a technique originally developed for interpretability[5]. LRP enables us to score and identify neurons' relevance to the main task. By pruning and retaining only features with high task relevance, we preserve the model's generalization and adaptability for fine-tuning to new subclasses and continual learning on new classes, as illustrated in Fig. 1.

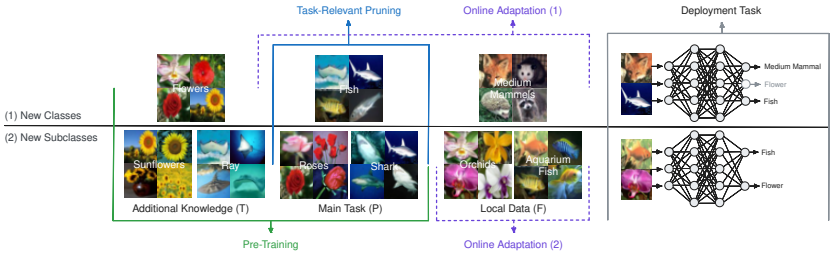


Fig. 1: Illustration of the two online adaptation scenarios on CIFAR-100.

In this paper, we first investigate how well a model compressed through structured pruning can adapt to new data. Then, we evaluate if our pre-training and task-relevant pruning method improves generalization during online adaptation on the CIFAR-10, CIFAR-100, and DomainNet datasets, comparing the results with both IMP and LRP-based pruning methods. Our experiments confirm that this approach can train a more accurate compressed model while achieving better generalization in both adaptation scenarios.

2 Background & Related Work

Layer-wise Relevance Propagation (LRP) [5] was first proposed to explain the predictions of neural networks by attributing relevance scores to features in the input data. LRP works by propagating a network's output back through the layers and assigning partial prediction contributions to each neuron based on its activation strength. The relevance score R_j^l of neuron j in layer l is computed (for LRP-0) by summing the relevance scores of all neurons in $l+1$ connected to j, weighted by their connection strength for the given input.

$$R_j = \sum_{k \in l+1} \frac{a_j w_{jk}}{\sum_{i \in l} a_i w_{ik}} R_k \tag{1}$$

Fine-tuning pruned models is often used to recover accuracy after pruning or as part of progressive pruning [3]. This is typically done with the original dataset to store behavior rather than adapt to new data. Gordon et al.[4] evaluated unstructured IMP on BERT before and after fine-tuning for transfer learning, finding that low pruning levels (30-40%) did not affect downstream tasks, and pruning once after pre-training was as effective as after fine-tuning to each task. Instead, we focus on structured pruning, which removes entire neurons or filters for higher efficiency but lower expressiveness at the same size.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

3 Methodology

Our proposed method, trained with extra knowledge (left in Fig. 2) learns more general features that are retained after pruning than a baseline trained only on the main task (right), The baseline is prone to memorize more task-specific features from the limited data, which are less relevant during online adaptation. By using knowledge-based pruning with only relevant task data, we effectively identify and remove redundant features while preserving model generalization. Magnitude-based pruning can only identify non-encoding neurons, rather than task-irrelevant ones. Despite requiring more active neurons during training, task-relevant knowledge-based pruning can still obtain high compression rates.

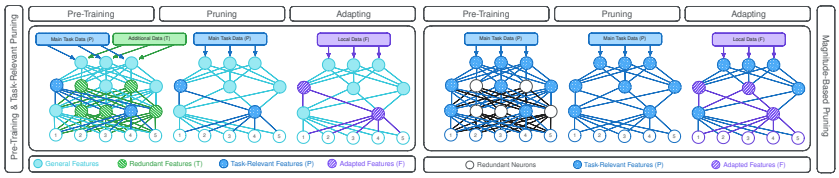


Fig. 2: The proposed pre-training and task-relevant pruning method, aimed at enhancing the generalization of compressed models for online adaptation.

We consider two different online adaptation scenarios:
Fine-tuning on new subclasses: the same classes are present during training and deployment, but the model must adapt to a shift in data distribution after pruning. We further divide the classes into distinct sets of subclasses, reserving some for fine-tuning and additional pre-training.
Continual learning on new classes: the model is trained on a subset of classes present during deployment. The remaining classes are only introduced after pruning. During training, some extra classes are included, but they are removed during pruning and not present at deployment.
Formally, the data is split into the following sets at the class or subclass level, depending on the scenario. These are then used as input for Algorithm 1.

Algorithm 1: Pre-Training and Task-Relevant Pruning

$M_0 \leftarrow \text{train}(T \cup P)$; `// Train on additional and task knowledge`
for $i \leftarrow 0$ **to** *prune_iterations* **do**
  $S_i \leftarrow$ initialize the list of relevance scores with zeros;
  for $d \in P$ **do**
    $a_d \leftarrow \text{forward}(M_i, d)$; `// Compute activations for input` d
    $S_d \leftarrow LRP(M_i, a)$ using equation 1;
    $S \leftarrow S + S_d$;
  $n_i \leftarrow$ compute l_1 norm of neurons/channels in S;
  $M_{i+1} \leftarrow$ remove neurons/channels from M_i with lowest % l_1 norms;
  $M_{i+1} \leftarrow \text{train}(M_{i+1}, P)$; `// Recover accuracy after pruning`

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

T: The extra data that is only used during the pre-training phase.
P: Data used during training and pruning to compute the LRP relevance scores.
F: A disjoint subset of data that is withheld for the fine-tuning phase.

Our model is pre-trained on $T \cup P$, pruned to retain only essential knowledge for classifying P, and adapted on F (*New Subclasses*) or $P \cup F$ (*New Classes*), depending on the scenario. To prevent forgetting original classes in the latter scenario, we use replay-based continual learning with the complete P dataset. We compared this to a baseline with $T = \varnothing$ to validate the impact of extra training knowledge on generalization. Additionally, we verify that LRP-based task-relevant pruning more effectively compresses the model while retraining all knowledge for classifying P, compared to IMP.

4 Experimental Setup

We use the CIFAR-10, CIFAR-100, and DomainNet datasets to evaluate our method. In CIFAR-10, we group the 10 classes into the superset {vehicles, animals}. For CIFAR-100, there are 20 superclasses containing 5 classes each. When fine-tuning, we reduce the number of classes to 2 and 20 respectively, distributing the original classes among T, P, and F. In continual learning, a distinct set of classes is reserved for T, P, and F, while maintaining the full 10 or 100 network outputs. We focus on fine-tuning to new styles for DomainNet, as it was designed specifically for domain adaptation, with the same 345 classes in 6 different styles. Table 1 contains the resulting splits, grouped in sets of two: one with additional training knowledge and one where $T = \varnothing$, as shown in Table 2. Splits for CIFAR-100 are analogous, but with too many classes to present here.

Name	T	P	F
CIFAR-9/3	birds, cats, deer, dogs, frogs, horses	airplanes, automobiles, trucks	ships
CIFAR-7/4	cars, dogs, horses	airplanes, birds, frogs, ships	cats, deer, trucks
DomainNet5/3	real, sketch	infograph, clipart, quickdraw	painting

Table 1: Overview of dataset splits for CIFAR-10 and DomainNet.

Scenario	New Classes		New Subclasses	
Pre-train on	$T \cup P$	$P \quad (T = \varnothing)$	$T \cup P$	$P \quad (T = \varnothing)$
CIFAR-10	CIFAR-9	CIFAR-3	CIFAR-7	CIFAR-4
CIFAR-100	CIFAR-90	CIFAR-30	CIFAR-60	CIFAR-40
DomainNet	N.A.	N.A.	DomainNet5	DomainNet3

Table 2: Overview of our experiment configurations.

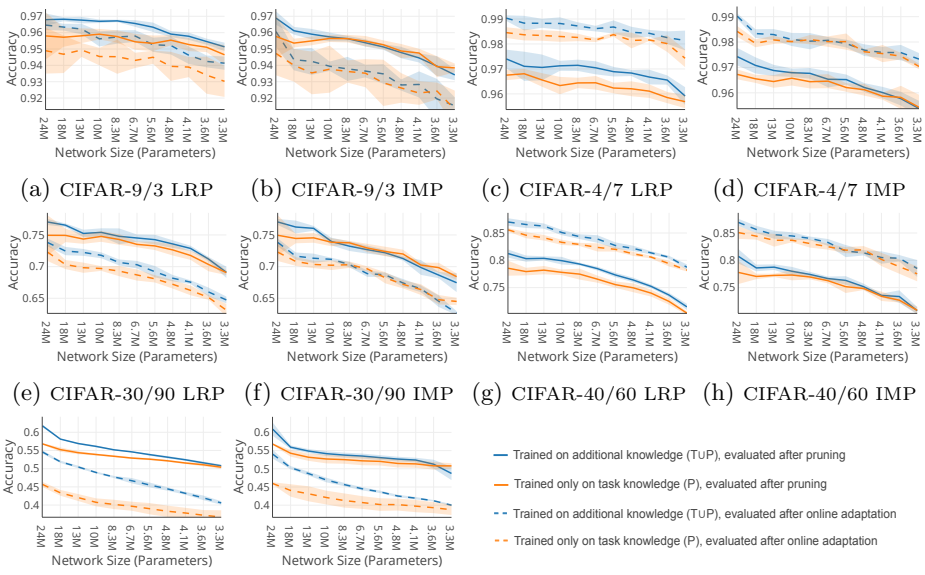
Experiments are repeated 5 times for each pruning method (LRP & IMP) to ensure consistent results. For IMP, weight magnitudes are used instead of relevance scores when computing the l_1 norm in Algorithm 1. Our ResNet-50 base architecture with 23.5M parameters is trained with a learning rate of 5×10^{-4}, batch size of 32, Adam optimizer, and an early stopping criterion based on validation

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

accuracy and patience of 4 epochs. We prune for 10 iterations, each removing 20% of the neurons in each layer, resulting in 17.5M, 13.3M, 10.3M, 8.3M, 6.7M, 5.6M, 4.8M, 4.1M, 3.6M, and 3.3M parameters. The checkpoint with the highest validation accuracy of each iteration is used in the next iteration, and during the adaptation phase to study the impact of model size on online adaptability.

5 Results

In this section, we assess the optimization of models on new data after either IMP or LRP-based task-relevant pruning, both with and without pre-training on additional knowledge. Fig. 3, shows that validation accuracy after online adaptation is consistently higher when the models are pre-trained on additional knowledge to improve generalization. The first point (23.5M) represents the full model accuracy without pruning, on P, and after online adaptation on F or $P \cup F$ depending on the scenario. Even the full model shows significant validation accuracy improvement from pre-training on additional knowledge, due to better generalization. This improvement is maintained successfully after our LRP-based pruning, with a consistent gap between the two pre-training approach accuracies on both P and F. However, this gap narrows as higher compression prunes more generalizable features, particularly for DomainNet (Fig. 3i).



(a) CIFAR-9/3 LRP (b) CIFAR-9/3 IMP (c) CIFAR-4/7 LRP (d) CIFAR-4/7 IMP

(e) CIFAR-30/90 LRP (f) CIFAR-30/90 IMP (g) CIFAR-40/60 LRP (h) CIFAR-40/60 IMP

(i) DomainNet3/5 LRP (j) DomainNet3/5 IMP

Fig. 3: Pruning and validation accuracy for all datasets and adaptation scenarios.

The IMP experiments only show a similar improvement in the first pruning iterations, as it cannot differentiate between task-relevant and additional features. This results in more retraining on P to recover the accuracy after pruning, over-

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

riding the general features learned in pre-training. We also conclude the models maintain reasonably high accuracy after pruning and adapt well to new data. In continual learning, $F \cup P$ accuracy is lower than on P after pruning due to the classification task being harder at the same capacity. For the CIFAR-10 and CIFAR-100 datasets, accuracy after fine-tuning is higher than post-pruning (Fig. 3c, 3d, 3g, 3h), as F contains fewer subclasses, simplifying the task. DomainNet experiments reveal both the largest accuracy drop and biggest benefit from pre-training on additional knowledge because the style in F (painting) is closer to T (real, sketch) than P (infograph, clipart, quickdraw). Heavy pruning and lack of general knowledge make adaptation harder for DomainNet3 which is optimized for abstract styles. In the last two IMP iterations with DomainNet5, accuracy on P drops below Domainet3 due to less intelligent pruning, although adaptation performance remains higher. These findings suggest that our pre-training and task-relevant pruning approach more effectively compresses models while preserving their generalization capabilities for online adaptation.

6 Conclusion

Adapting to new data is crucial for edge devices in dynamic environments, but their limited resources restrict model size and complexity. This necessitates model compression such as pruning, but these typically do not account for online learning scenarios. We present a novel approach that extends the generalization of pruned models for online adaptation without compromising compression rates. By pre-training on additional knowledge and using LRP to compute relevance scores, we identified and retained only neurons encoding task-relevant knowledge during pruning, while still benefiting from the increased generalization. Experiments on CIFAR-10, CIFAR-100, and DomainNet confirmed that this increased generalization resulted in higher validation accuracies, which were better maintained post-pruning compared to an IMP baseline. The improved generalization significantly enhanced accuracy during online adaptation, both for learning new classes and when fine-tuning to new subclasses. Our approach allows for effective model compression while maintaining high generalization on new data, making it suitable for online adaptation scenarios.

References

[1] Xu et al. 'Rethinking Network Pruning - under the Pre-Train and Fine-Tune Paradigm'. CoRR, vol. abs/2104.08682, 2021.

[2] Kim et al. 'A Broad Study of Pre-Training for Domain Generalization and Adaptation'. CoRR, vol. abs/2203.11819, 2022.

[3] Blalock, Davis W., et al. 'What Is the State of Neural Network Pruning?' CoRR, vol. abs/2003.03033, 2020.

[4] Gordon, Mitchell A., et al. 'Compressing BERT: Studying the Effects of Weight Pruning on Transfer Learning'. CoRR, vol. abs/2002.08307, 2020.

[5] Bach et al. 'On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation'. PLOS ONE, vol. 10, no. 7, Public Library of Science, 07 2015, pp. 1–46.